Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Comparative Study of Feature Extraction and Classification Techniques for Punjabi Speech Recognition

Jaspreet Kaur Sandhu [1], Munish Kumar [1], AMITOJ SINGH [2]
[1] *Department of Computational Sciences, Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India*
[2] *School of Sciences and Emerging Technologies, Jagat Guru Nanak Dev Punjab State Open University, Patiala, Punjab, India*

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Punjabi is the second most spoken language in north India. It is desirable to have a communication system in a local language that permits ordinary people to communicate with machines via speech interface to retrieve information or to perform their daily activities. It is observed that conventional Automatic Speech Recognition (ASR) systems are in English or European languages. They also use the Mel Frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction (PLP) etc. features, but do not perform well in real-world situations. Here, a study has been carried out on different feature extraction techniques to find the best-performing feature extraction technique for Punjabi language in clean and noisy environments. This paper also compares the performance of different acoustic models-based ASR systems for Punjabi. Previously performed studies have utilized the Context-Independent (CI) and Context-Dependent (CD) acoustic models but this study focused on CD models. This study will help the researcher to know about the behavior of different feature extraction techniques and acoustic models for Punjabi speech dataset in clean and noisy environments. Experimental results show that MFCC and Gammatone Frequency Cepstral Coefficients (GFCCs) perform well in clean and noisy environments, respectively. The best Word Error Rate (WER) is 12% and 14.8% achieved by MFCC and GFCC feature extraction technique with Bidirectional Long Short-Term Memory (BLSTM) as acoustic model in clean and noisy environment, respectively.

Keywords: Acoustic Model, BLSTM, CNN, GFCC, Feature Extraction Techniques.

INTRODUCTION

In human communication, speech plays an important role. Therefore, various languages are spoken in the world by human beings for communication. A computer system that understands the spoken language can be very useful in various areas like agriculture, healthcare, and government sectors, etc. ASR is the ability of the machine to translate spoken words into written form (Pasricha & Aggarwal, 2016). Although the last decades have witnessed significant progress in ASR. Still, in many real usage scenarios, the performance of ASR systems is lagging far behind human-level performance because quasi-stationary nature of speech signals.

A speech recognition system consists of five blocks: - Feature extraction, Acoustic modeling, Pronunciation modeling, Language modeling, and Decoder. Feature extraction is the most important phase in a speech recognition system. ASR faces some problems during the feature extraction process because of the variability of the speakers (Guglani & Mishra, 2020). During feature extraction, the speech signal is converted into discrete sequence of feature vectors, which is assumed to contain only the relevant information about given utterance that is important for its correct recognition. An important property of feature extraction is the suppression of irrelevant information for correct classification such as information about speaker (e.g. fundamental frequency) and information about transmission channel (e.g. characteristic of a microphone). However, the information conveyed by these feature vectors may be correlated and less discriminative which may slow down the further processing. Feature extraction methods like MFCC provide some way to get uncorrelated vectors through Discrete Cosine Transform (DCT). Many new feature extraction techniques have led to significant advances in ASR.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Acoustic modeling is an open research domain in ASR. Almost all the traditional ASR systems are Hidden Markov Model (HMM) based. The relationship between HMM states and the acoustic input is usually represented by Gaussian Mixture Models (GMMs) or Artificial Neural Networks (ANNs). However, the ANNs were typically trained only with one hidden layer. Earlier, it was suspected that deep networks could model complex higher statistical structures effectively (Mohamed et al., 2011). Many researches indicated that Deep Neural Network(DNN) based acoustic models can outperform GMMs in many speech recognition tasks (T. N. Sainath, Kingsbury, Soltau, & Ramabhadran, 2013). As first introduced in (Dahl, Yu, Deng, & Acero, 2012), the CI pre-trained DNN/HMM hybrid architectures have been proposed for phone recognition. Then, CD pre-trained DNN/HMM for large vocabulary speech recognition is studied and discussed in (Yu, Seide, & Li, 2012) and achieved very competitive performance, and have become the focus of ASR research.

The introduction of DNNs-based acoustic models changed many conclusions based on GMMs, owing to the difference that neural network is a discriminative model and the other is generative model. Section 3 of this paper focuses on the choice of acoustic models in ASR. The conducted experiments demonstrate the performance of different acoustic models for ASR.

Over the past years, several review papers were published, in which the ASR task was examined from various perspectives. This review discusses some of the ASR challenges and also presents a brief overview of number of well-known ASR systems methodologies. The authors consider various feature extraction techniques: MFCC, PLP, GFCC, Linear Predictive Coding (LPC), Linear Prediction Cepstral Coefficients (LPCCs), Relative Spectral Transform (RASTA), Power Normalized Cepstral Coefficients (PNCCs) as well as eight different acoustic models: HMM, GMMs, DNNs, Convolutional Neural Networks (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), BLSTM, Gated Recurrent Units (GRUs).Finally, the performance comparison is presented based on the feature extraction and classification techniques used. In this paper, two research questions are tried to answer for future Punjabi ASR research.

RQ1: What is the best feature extraction technique among different feature extraction techniques used in Punjabi ASR?

RQ2: Which is the best acoustic model among the different acoustic models used in Punjabi ASR?

This paper is organized as follows. The next section, i.e., section 2 presents the popular and widely used various feature extraction techniques. Section 3 describes the various acoustic models in detail, while section 4 introduces the details of Punjabi speech corpus, experimental setup used for this study, experimental results and analysis etc. Section 5 concludes the paper with some discussions.

FEATURE EXTRACTION TECHNIQUES

Feature extraction is a crucial step in ASR where an audio signal is transformed into a representation, known as features, which is used by classifiers for recognition tasks. After decades, feature extraction is still an open field of research in ASR field. Here, fundamental techniques focus on extracting frequency information from unprocessed audio files to filter out unwanted disruptive sounds and reverberation to influence the speech recognition system. Additionally, certain approaches explore the effect of temporal information extracted from speech. Some common feature extraction techniques used in ASR tasks are: MFCC (Davis & Mermelstein, 1990), Linear Predictive Coding (LPC) (O'Shaughnessy, 1988), PLP (Hermansky, 1990), RASTA (Hermansky & Morgan, 1994; Koehler, Morgan, Hermansky, Hirsch, & Tong, 1994), gammatone filterbank features (Zhao & Wang, 2013), and spectral contrast.

Mel-Frequency Cepstral Coefficients (MFCCs)

MFCC is the most widely used feature extraction technique in ASR systems (Davis & Mermelstein, 1990). MFCC feature is considered the dominant characteristic parameter that is based on the human being auditory system. The working mechanism of MFCCs is a duplication of the human being auditory system. It is derived from the power spectrum of the audio signal after implementing several stages. Pre-emphasis is the first step where the suppressed frequency component is compensated by passing through the filter of high pass. The second step is framing, applied on a continuous signal to make it discrete by diving into the frames. To maintain the continuity between succeeding frames, the Hamming window is applied. After this, Fast Fourier Transformation (FFT) is performed to convert the

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

discrete time domain into the frequency domain. Then, computation of signal energy is done in various frequency band in the filter bank processing. To do this, Davis and Mermelstein (Davis & Mermelstein, 1990) developed the Mel scale (given in eqn. 1) that makes use of triangular filters for wrapping the magnitude spectrum.

$$Mel(f) = 2595\log_{10}\left(1 + {}^{f}/_{700}\right) \quad (1)$$

The speech vectors are made significant by calculating the logarithm of the square magnitude of the coefficients and cepstral domain coefficients are obtained by the DCT and result in final MFCC features. The block diagram depicting the steps needed to compute MFCC is given in Figure 1.

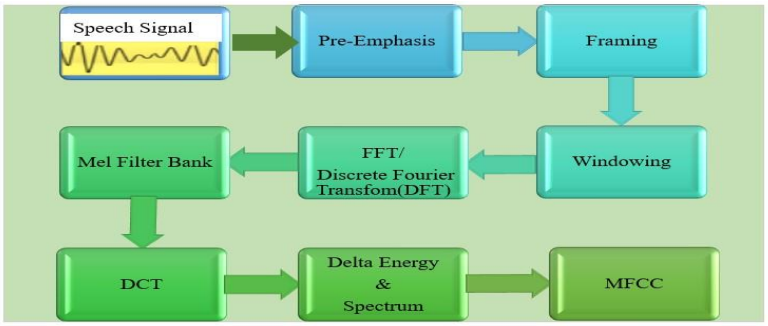


Figure 1:Block diagram of MFCC

Perceptual Linear Prediction (PLP)

PLP is similar to MFCCs but incorporates aspects of human hearing perception in its feature extraction process. It uses a non-linear frequency scale to better match human hearing perception and applies linear prediction to the audio signal before passing it through the bark filter bank. The block diagram illustrating the steps in computing the PLP is given in Figure 2. The description of PLP processing is given below:

The initial few steps, i.e., preprocessing, framing, windowing, and fast Fourier transformation are almost the same as already discussed in MFCC. A Bark-scale filter bank divides the frequency spectrum of an audio signal into bands according to the Bark scale (Hermansky, 1990). This allows for a more perceptually relevant signal representation compared to a linearly spaced filter bank. Equal loudness pre-emphasis contours represent the relationship between sound intensity and perceived loudness at various frequencies. In the next step, intensity loudness conversion is done through a cubic-root amplitude compression. Then, autoregressive modeling and coefficient calculation are made to get the coefficients. The resulting autoregressive coefficients (frequency axis) are transformed into the bark scale. It uses a bark scale to work on the principle of human hearing resolution in frequency. PLP analysis is computationally efficient and yields a low-dimensional representation of speech.

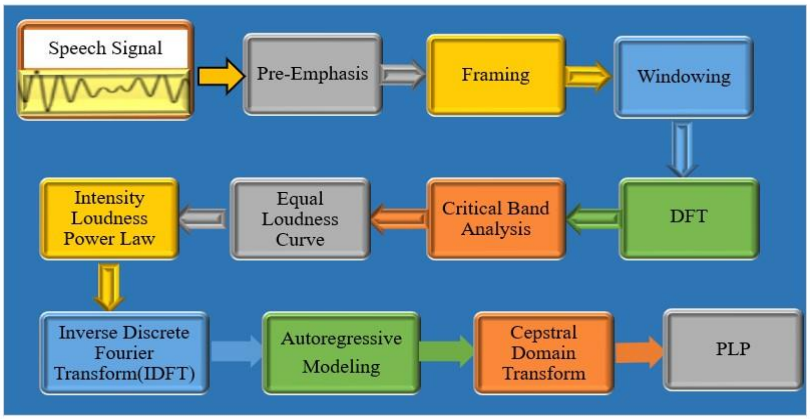


Figure 2: Block diagram of PLP

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Gammatone Frequency Cepstral Coefficients (GFCC)

GFCCs are features derived from the auditory filter responses in the Gammatone filter bank (Zhao & Wang, 2013). These coefficients capture the energy distribution across different frequency bands. The coefficients are computed by convolving the input signal with a set of Gammatone filters and extracting the resulting filterbank energies. Its initial steps are the same as MFCC, i.e., Pre-emphasis, framing, windowing, and fast Fourier transformation. Mel filterbank is replaced with Gammatone filterbank, rest of the steps are also like MFCC. Gammatone frequency spectral coefficients are widely used in speech and audio processing tasks, including and speaker identification. They offer better discrimination between different sound sources and are robust to noise. Gammatone-based features have been shown to outperform traditional filterbank-based features in challenging acoustic environments. In Figure 3, the steps involved in the calculation of GFCC is listed.

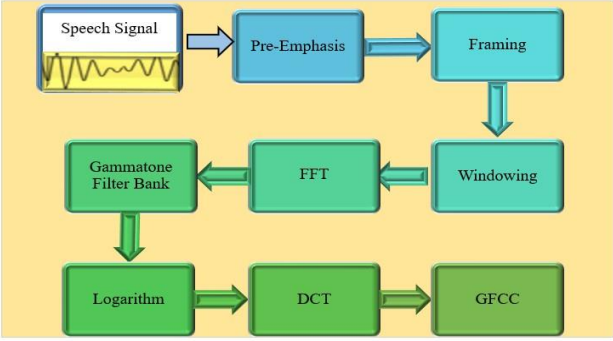


Figure 3: Block Diagram of GFCC Computation

Linear Predictive Coding (LPC)

LPC is a technique used in speech recognition to model the spectral envelope of a speech signal (O'Shaughnessy, 1988). It assumes that speech is produced by a time-varying linear filter excited by a source signal. The technique involves estimating the parameters of this linear filter to capture the characteristics of the speech signal. LPC models the speech signal as the output of a linear prediction filter, typically using an autoregressive model. Every frame of the windowed signal is auto-correlated, whereas the order of the linear prediction analysis is the highest value of autocorrelation. For estimating LPC coefficient, Yule-Walker equation is used in which the autocorrelation function is utilized. By estimating the coefficients of this filter, LPC captures the formant frequencies and spectral shape of the speech signal. LPC coefficients can be used for speech analysis, synthesis, and compression. They provide compact representations of the speech signal for classification and identification purposes.

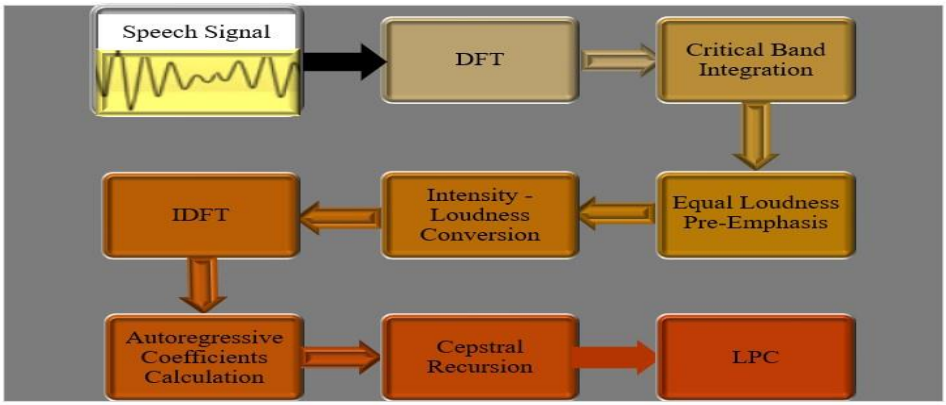


Figure 4: Computation of LPC

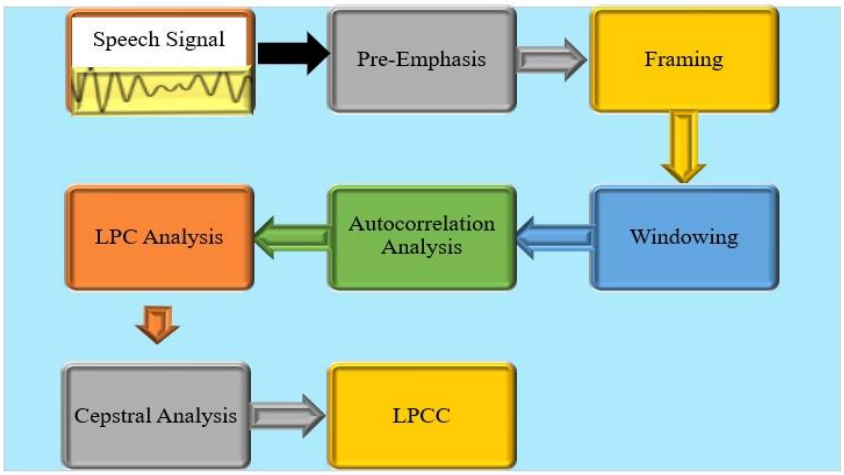
Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Linear Prediction Cepstral Coefficients (LPCCs)

LPCCs combine LPC with cepstral analysis to extract features from speech signals (Quen-Zong, Jou, & Suh-Yin, 1997). They are obtained by taking the cepstral coefficients of the LPC residual signal. They capture both spectral and temporal characteristics of speech, making them effective for speech recognition tasks. LPCCs are computed by taking the DCT of the log of the power spectrum of the LPC residual signal. LPCCs are robust to additive noise and channel distortion due to their compact representation of the speech signal. LPCCs offer a more compact representation compared to MFCCs while maintaining discriminative power. Steps for calculating LPCC is listed in Figure 5.



.

Figure 5: Computation of LPCC

Relative Spectral Filtering (RASTA)

RASTA is a technique used in speech recognition to enhance robustness against various acoustic disturbances (Hermansky & Morgan, 1994). It operates by smoothing the short-term spectral changes of speech signals while preserving long-term spectral characteristics. RASTA filtering reduces the impact of channel distortions, noise, and other environmental factors on speech signals. It achieves this by applying a high-pass filter that emphasizes slow spectral changes and attenuates rapid fluctuations. RASTA filtering helps in mitigating the effects of reverberation and other time-varying distortions, making speech signals more intelligible. It has been applied to improve the performance of ASR systems in noisy conditions. RASTA filtering is particularly effective in scenarios where speech signals are degraded by reverberation or background noise.

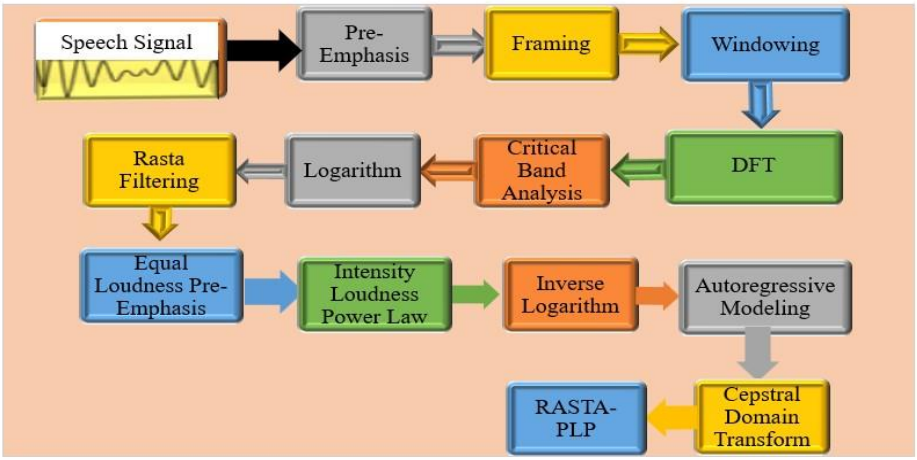


Figure 6: Block diagram of RASTA filtering included with PLP (RASTA-PLP)

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

It is used to improve speech quality in noisy environments and reduce noise impact. The high pass filtering in RASTA helps in alleviating the convolutional noise effects that occurs in the channel, whereas with the low pass filtering the frame smoothing is done for framing the spectral changes. It can be used directly as features or combined with other techniques such as PLP (Koehler et al., 1994).

Power Normalized Cepstral Coefficients (PNCCs)

PNCCs are a type of feature representation used in ASR systems (Kim & Stern, 2016). The main characteristic of PNCC is the application of power normalization to the filterbank energies before computing the cepstral coefficients, which helps in improving the robustness of the features to varying input signal power levels. PNCCs offer several advantages over traditional MFCCs, such as improved noise robustness, better discrimination between speech and non-speech sounds, and reduced sensitivity to channel effects.

The computation of PNCC involves applying a power-law compression to the filterbank energies to simulate the human auditory system's non-linear response to sound intensity. This normalization helps in emphasizing important spectral features while suppressing noise and irrelevant information. PNCCs have been shown to outperform MFCCs in challenging acoustic conditions, including noisy environments and reverberant conditions. PNCCs provide a robust and discriminative representation of speech signals, contributing to the improvement of ASR system performance in real-world scenarios. In Figure 7, the steps for computing PNCC are shown.

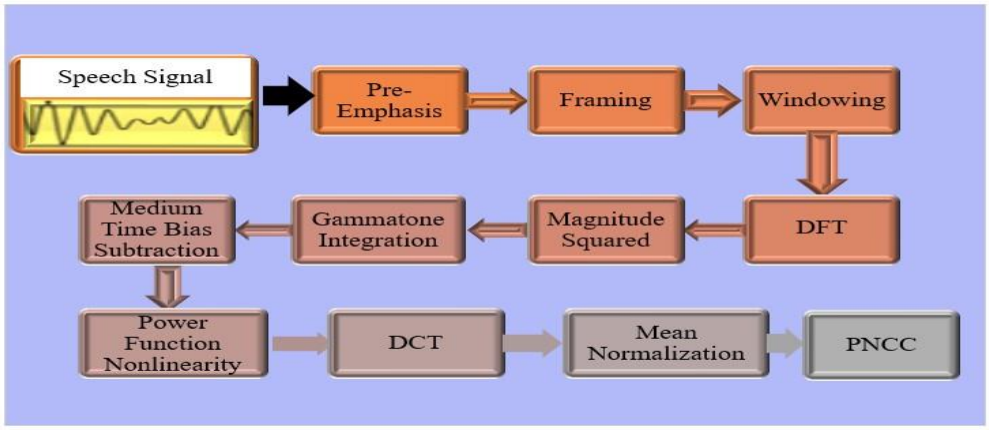


Figure 7: Computation of PNCC

The above discussed techniques have its own advantages and is suitable for different applications and scenarios in ASR. A combination of these techniques is also used to extract a comprehensive set of features for training ASR systems. Table 1 shows the advantages and disadvantages of each feature extraction technique.

Table 1: advantages and disadvantages of feature extraction techniques

Features	Advantages	Disadvantages
MFCC	• Representation of Human Auditory Perception. • Dimensionality Reduction. • Useful in Genre Classification • Widely Used	• Sensitivity to Noise and Variability. • Limited Temporal Resolution • Loss of Phase Information
PLP	• Human Hearing Perception • Noise Robustness • Dimensionality Reduction	• Complexity • Tuning Parameters • Limited Application Scope
GFCC	• Robustness to Noise • Improved Speaker and Speech Recognition • Frequency Resolution	• Computational Complexity • Resource Intensive
LPC	• Efficient Compression.	• Sensitivity to Noise

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

	• Mathematical Simplicity • Real-Time Processing	• Voiced/Unvoiced Detection • Phase Information Loss
LPCC	• Compact Representation • Good Discrimination • Stable and Robust Features	• Sensitivity to Noise • Loss of Phase Information • Limited Effectiveness for Non-Speech Signals
RASTA	• Robustness to Channel Distortions • Temporal Filtering • Compatibility with Other features	• Sensitivity to Speech Variability • Parameter Tuning • Limited to Speech Signals
PNCC	• Non-linear Power Law • Temporal Masking • Resilience to Variability	• Difficult Implementation • Limited Adoption • Parameter Sensitivity

CLASSIFICATION TECHNIQUES

The acoustic model's primary task is to represent the relationship between speech signals and linguistic units, such as phonemes or sub-word units. Various acoustic models are found well-suited for this task because they can model the temporal dependencies in speech signals effectively. From the start of the ASR process, many acoustic models have been proposed and successfully applied. Some popular acoustic models have been discussed in this section.

Hidden Markov Model (HMM)

HMM is widely used as an acoustic model in ASR. It is a statistical model developed by Baum in 1967 (Baum & Eagon, 1967). Speech signals are usually pre-processed to extract relevant acoustic features such as MFCCs, which represent the spectral characteristics of the speech signal over time. In an HMM, each phoneme or sub-word unit is represented by a state, and the transitions between states are governed by probabilities. The states are hidden because they cannot be directly observed; instead, the acoustic features associated with each state are observed.

Each state in the HMM emits a probability distribution over the observed acoustic features. These emission probabilities represent how likely it is to observe certain acoustic features given the state of the HMM. The parameters of the HMM, including transition probabilities and emission probabilities, are estimated from a large corpus of labeled speech data. This training process typically involves algorithms like the Baum-Welch algorithm (Baum, Petrie, Soules, & Weiss, 1970) or Maximum Likelihood Estimation (Myung, 2003). During decoding, the HMM is used to find the most likely sequence of states given the observed acoustic features. This is typically done using the Viterbi algorithm, which efficiently finds the most probable state sequence. The output of the acoustic model is combined with a language model to produce the final recognition result.

Overall, hidden Markov models have been a milestone of acoustic modeling in speech recognition for several decades, although recent approaches, such as DNNs and RNNs, have gained popularity due to their ability to capture more complex patterns in the data. However, HMMs still play a vital role in traditional ASR systems, especially in combination with other techniques.

Gaussian Mixture Models (GMMs)

GMM is used as an acoustic model in statistical speech recognition systems. GMM is trained using the extracted features from speech signals. The training data typically consists of pairs of feature vectors and their corresponding phonetic labels. The GMM is trained to model the distribution of features corresponding to each phoneme. Each phoneme is represented by a separate GMM. This allows the model to capture the variability in the acoustic characteristics of different phonemes. Each GMM is composed of multiple Gaussian components, each representing a cluster of feature vectors in the high-dimensional feature space. These components are jointly trained to capture the statistical properties of the corresponding phoneme or sub-word unit.

During recognition, an input speech segment is given, the likelihood of the observed features is computed for each phoneme. This is done by evaluating the probability density function of the GMM for the observed features. The

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

likelihood scores obtained from the GMMs are then used as input to a decoder, which combines them with other linguistic models (such as HMM) to generate the most probable sequence of phonemes. GMMs still serve as a fundamental baseline and are used in hybrid systems where they are combined with HMM for improved performance.

Deep Neural Networks (DNNs)

In recent years, DNNs have become a dominant approach for acoustic modeling in ASR systems (Dahl et al., 2012; Hinton et al., 2012). Like traditional acoustic models, it processes the acoustic features extracted from the raw audio signal. DNNs consist of multiple layers of neurons organized in a feedforward manner. These networks often include several hidden layers, allowing them to learn complex patterns in the input data.

DNNs are trained using supervised learning techniques, where they are presented with pairs of input acoustic features and corresponding target labels (e.g., phoneme or word sequences). The network's parameters (weights and biases) are fixed iteratively to minimize a loss function, such as Cross-Entropy (CE) (T. N. Sainath et al., 2013) or Mean Squared Error (MSE) (Chen, Xing, Liang, Zheng, & Príncipe, 2014), which measures the difference between the predicted output and the true labels. DNN learns to map input acoustic features to output labels by iteratively adjusting its parameters during training process. The network automatically learns to extract relevant features from the input data and capture complex relationships between the input features and output labels.

DNNs can also be used to model context-dependent phonetic units, such as triphones, which capture the phonetic context of each speech segment. This allows the model to better capture the variability in speech sounds depending on their surrounding context. The output of the DNN acoustic model is typically combined with a language model during the decoding process to produce the final recognition result.

DNN-based acoustic models have shown significant improvements in ASR performance compared to traditional approaches such as HMMs, particularly when trained on large amounts of data. They are also capable of capturing complex patterns in the data.

Convolutional Neural Network (CNN)

One of the popular kinds of deep learning architecture is CNN, which is widely used in vision tasks (Ren & Xu, 2015). CNN has changed the paradigm of ASR as an acoustic model (Passricha & Aggarwal, 2019). The various attractive advancements of CNN are weight sharing, pooling, and convolutional filters. CNNs have gained popularity as acoustic models in ASR tasks, especially for processing raw audio waveforms directly (Palaz, Magimai-Doss, & Collobert, 2019).

The building blocks of CNNs is convolutional layers. These layers consist of filters, also known as kernels, that slide over the input data, performing convolutions to extract local features. In the context of acoustic modeling, these filters capture various aspects of the input audio signals, such as temporal patterns and frequency content. After each convolutional layer, pooling layers are often used to reduce the spatial dimensionality of the feature maps, while retaining the most important information. Max pooling is a common technique where the maximum value within each pooling window is retained, effectively down-sampling the feature maps (Passricha & Aggarwal, 2020).

CNNs consist of multiple convolutional layers stacked on top of each other, allowing the network to learn hierarchical representations of the input data. Each successive layer captures increasingly abstract features, potentially representing higher-level acoustic characteristics. In last, one or more fully connected layers may be added to the network. These layers integrate the extracted features from the convolutional layers and perform classification tasks, depending on the specific ASR architecture.

CNNs are trained using supervised learning techniques, where they are presented with audio waveforms and corresponding target labels. The network's parameters are adjusted iteratively to minimize a loss function, such as cross-entropy, which measures the discrepancy between the predicted output and the true labels.

Unlike traditional acoustic models, which often use handcrafted features like MFCCs, CNNs can directly process raw audio waveforms as input. The raw waveform is typically divided into short-time segments, such as frames of 20-50 milliseconds, with some overlap (Palaz et al., 2019).

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Research Article

CNN-based acoustic models offer several advantages, including the ability to learn hierarchical representations directly from raw audio data, scalability to large datasets, and effectiveness in capturing local and global patterns in the input signal. They have demonstrated state-of-the-art performance in various ASR tasks and are widely used in modern speech recognition systems.

Recurrent Neural Networks (RNNs)

RNNs are another powerful type of neural network architecture used as an acoustic model (Graves, Mohamed, & Hinton, 2013). Unlike DNNs or CNNs, RNNs are designed to handle sequential data, making them well-suited for processing time-series data such as speech signals.

RNNs process sequential data in one-time step at a time, where each time step corresponds to a frame of the input signal. This allows the network to capture temporal dependencies and patterns in the input sequence. The key characteristic of RNNs is their recurrent connections, which allow information to persist over time. At each time step, the network receives input from the current frame as well as information propagated from the previous time step through the recurrent connections. RNNs maintain a hidden state vector that represents the network's internal memory or context at each time step. This hidden state is updated recursively based on the current input and the previous hidden state, allowing the network to capture temporal dynamics in the input sequence (Robinson, Hochberg, & Renals, 1996).

RNNs are trained using supervised learning techniques, where they are presented with audio waveforms and corresponding target labels. The network's parameters are adjusted iteratively to minimize a loss function, such as cross-entropy, which measures the differences between the actual outputs and the targeted outputs. Gradient descent-based optimization algorithms, such as backpropagation through time, are commonly used for training RNNs.

RNN-based acoustic models have shown promising results in ASR tasks, especially when trained on large amounts of labeled speech data. They excellently capture temporal dependencies in sequential data and have been widely adopted in various ASR applications.

Long Short-Term Memory (LSTM)

LSTM networks are an advanced type of RNN architecture that has been successfully used as an acoustic model (Hochreiter & Schmidhuber, 1997). LSTM networks are designed to process sequential data, making them well-suited for modeling the temporal dynamics of speech signals (Sak, Senior, & Beaufays, 2014). LSTMs contain memory cells that maintain information over time, allowing them to capture long-range dependencies in the input sequence. These memory cells have a self-connected recurrent structure, which enables them to selectively retain or discard information from previous time steps.

LSTMs use gating mechanisms to control the flow of information through the network and regulate the memory cell's state. These gating mechanisms include an input gate, a forget gate, and an output gate, each of which consists of a sigmoid activation function and element-wise multiplication operations. These gates enable the LSTM to learn when to update the memory cell's state and when to forget irrelevant information.

LSTM networks are trained using supervised learning techniques, where they are trained with pairs of input sequences and corresponding target labels. The network's parameters including those of the memory cells and gating mechanisms, are set iteratively to minimize a loss function, such as cross-entropy or mean squared error.

LSTM networks have several advantages as acoustic models for ASR. LSTMs are capable of capturing long-range dependencies in sequential data, making them effective for modeling speech signals, which often exhibit complex temporal patterns. LSTMs address the vanishing gradient problem commonly encountered in traditional RNNs, allowing them to learn from sequences with long durations more effectively. LSTMs require fewer parameters compared to other types of recurrent networks, making them computationally efficient and easier to train. Overall, LSTM networks have demonstrated strong performance as acoustic models and have been widely adopted in both research and industrial applications (Li & Wu, 2015).

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Bidirectional Long Short-Term Memory (BLSTM)

BLSTM networks are a variant of RNNs. BLSTMs combine the advantages of LSTMs with bidirectional processing, allowing them to capture both past and future contexts in the input sequence (Passricha & Aggarwal Rajesh, 2019). BLSTMs process the input sequence in both forward and backward directions simultaneously. This means that at each time step, the network receives information from both past and future contexts, enabling it to capture temporal dependencies in both directions.

A BLSTM network consists of two sets of LSTM layers. One set processes the input sequence in the forward direction, while the other set processes it in the backward direction. Each set of LSTM layers maintains its own hidden state and memory cells. The outputs of the forward and backward LSTM layers at each time step are typically concatenated to create a combined representation of the input sequence. This combined representation captures the past and future context of each time step. BLSTM networks are generally trained in a supervised manner, like other neural networks.

BLSTMs offer several advantages as acoustic models for ASR. BLSTMs capture past and future contexts, allowing them to model long-range dependencies in the input sequence more effectively (Passricha & Aggarwal Rajesh, 2019). BLSTMs are robust to variations in the timing of speech events, as they can leverage information from both preceding and succeeding frames. By capturing bidirectional context, BLSTMs can provide richer representations of the input sequence, leading to improved recognition accuracy in ASR tasks. Overall, BLSTM networks have demonstrated strong performance as acoustic models in ASR systems and are commonly used in research and industrial applications.

Gated Recurrent Units (GRUs)

GRUs are another variant of RNNs that also used as acoustic models (Ravanelli, Brakel, Omologo, & Bengio, 2018). GRUs are similar to LSTM networks but have a simpler architecture with fewer parameters. Like other RNN architectures, GRUs process sequential data, making them suitable for modeling the temporal dynamics of speech signals. They also operate on input sequences one-time step at a time, where each time step corresponds to a frame of the input signal. GRUs use gating mechanisms to control the flow of information through the network. Unlike LSTMs, which have separate input and forget gates, GRUs have a single gate called the update gate. The update gate determines how much of the previous hidden state should be retained and how much of current input should be incorporated into the new hidden state.

In addition to the update gate, GRUs have a reset gate that controls how much of the previous hidden state should be reset before computing the new hidden state. The reset gate allows GRUs to selectively forget irrelevant information from the past. GRU networks are trained using supervised learning techniques. The network's parameters, including those of the update and reset gates, are fixed iteratively to minimize a loss function, such as cross-entropy or mean squared error.

GRUs offer several advantages as acoustic models for ASR. GRUs have a simpler architecture compared to LSTMs, which may lead to faster training and lower computational complexity. GRUs have fewer parameters as compared to LSTMs, making them more memory-efficient and easier to train, especially on smaller datasets. Despite their simpler structure, GRUs are capable of capturing long-range dependencies in sequential data, making them suitable for modeling speech signals. Overall, GRUs have shown promising results as acoustic models in ASR systems. They provide an effective and efficient alternative to more complex RNN architectures like LSTMs.

Table 2: advantages and disadvantages of different acoustic models

Acoustic Model	Advantages	Disadvantages
HMM	• Able to Model Time Distribution of Speech signal • Probabilistic Framework • Integration with Other Models	• Assumption of Independence • Fixed Model Topology • Large State Space
GMM	• Flexibility in Modeling data Distribution	• Sensitivity to Initialization

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

	• Soft Clustering • Robustness to Outliers	• Computational Complexity • Curse of Dimensionality
DNN	• High accuracy • Highly Adequate for Pattern Recognition Applications • Self-organising • Self-learning	• Computational Resources • Overfitting • Hyperparameter Tuning
CNN	• Able to Extract Features from Raw Data • Capture Spatial Hierarchies of Features • Parameter Sharing • Robust to Variations	• Computationally Intensive • Need Large Datasets • Complex architecture
RNN	• Sequential Data Handling • Memory of Previous Inputs • Context Modeling	• Vanishing and Exploding Gradients • Overfitting • Gradient Computation
LSTM	• Long-Term Dependency Handling • Ability to Learn Temporal Patterns • Gated Memory Units	• High Computational Cost • Difficulty in Training • Limited Interpretability
BLSTM	• Bi-directional Context • Improved Sequence Modeling • Reduced Information Loss	• High computational complexity • Potential Overfitting • Parallelization Challenges
GRU	• Simpler Architecture • Faster Convergence • Modeling Complex Patterns	• Limited Modeling Capability • Reduced Control over Memory • Potential Information Loss

EXPERIMENTS

This section discusses the speech corpus used, baseline architectures, hardware details, and the model description.

Punjabi Speech Corpus

The effectiveness of ASR systems relies on the accessibility of labeled speech data for training. Punjabi is categorized as an under-resourced language due to the scarcity of extensive speech corpora, benchmarked data, and other necessary resources. The ASR systems performance was evaluated using Punjabi speech corpus with tonal characteristics which included a total of 119,500 utterances recorded from 180 speakers out of which 106 were male and 74 were female belonging to the Malwa, Majha, Doaba, and Powadh regions of Indian Punjab. The corpus contained frequently used Punjabi words, Punjabi tonal words, phonetical rich Punjabi sentences, , affirmatives Punjabi sentences , doubtful Punjabi sentences ,interrogative Punjabi sentences and sentences having tonal words. The description of the Punjabi speech corpus with tonal characteristics is given in Table 3.

Table 3: Detail of Punjabi Speech Corpus with Tonal Characteristics

Speaker type	**Region**	**Male**	**Female**	**Words spoken**	**Time duration (hrs)**
Set1	Malwa	28	27	35,350	19:15
Set2	Majha	30	15	30,100	16:30
Set3	Doaba	30	20	34,300	18:30
Set4	Powadh	18	12	19,750	10:45
Total Speakers = 180			Total Utterances = 119,500		

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Baseline

The performance of different ASR systems is evaluated for seven feature extraction methods, i.e., MFCC, PLP, GFCC, LPC, LPCC, RASTA, PNCC. Figure 8 demonstrates the combined architecture of various feature extraction techniques used with different acoustic models. The systems are evaluated for clean as well as real-time conditions. The features containing the highest information about speech signals are kept to reduce the computational cost. First 13 coefficients with their first and second-order derivatives, i.e., 39 coefficients are used. To extract features, 25ms long hamming window is used with a consistent shift of 10ms. Cepstral Mean and Variance Normalization(CMVN) (Viikki & Laurila, 1998) is an efficient normalization technique which is adopted by almost every speech recognition system to increase the robustness. The same practice, i.e., normalized speech data with zero mean and unit variance, is followed for training and testing purposes. Decoding is defined as the task of recognizing the speech samples based on their acoustic characteristics. For decoding, a trigram language model is used. Fundamental dissimilarities between training data and generated output degrade the accuracy of ASR.

The HMM consists of a set of states representing different phonetic units (e.g., phonemes or sub-phonetic units). Each state has a GMM which represents the probability distribution of acoustic features emitted by that state. Each state emits a mixture of Gaussians representing the acoustic features associated with that state. More mixtures allow for more complex sound representations but increase model complexity. The HMM configuration is used in experiments, total 4 states per phoneme are used having 8 Gaussians per mixture. Also, left-to-right topology is used. Transition probabilities between states are typically uniform within a state level and small probabilities of transitioning to neighboring states are allowed for slight variations in timing. The probability of transitioning from one state to another is taken as 0.2. Covariance type used is diagonal. Moreover, emitting states across different phonemes are tied together to share the same GMM. It reduces the number of parameters to learn and improves efficiency.

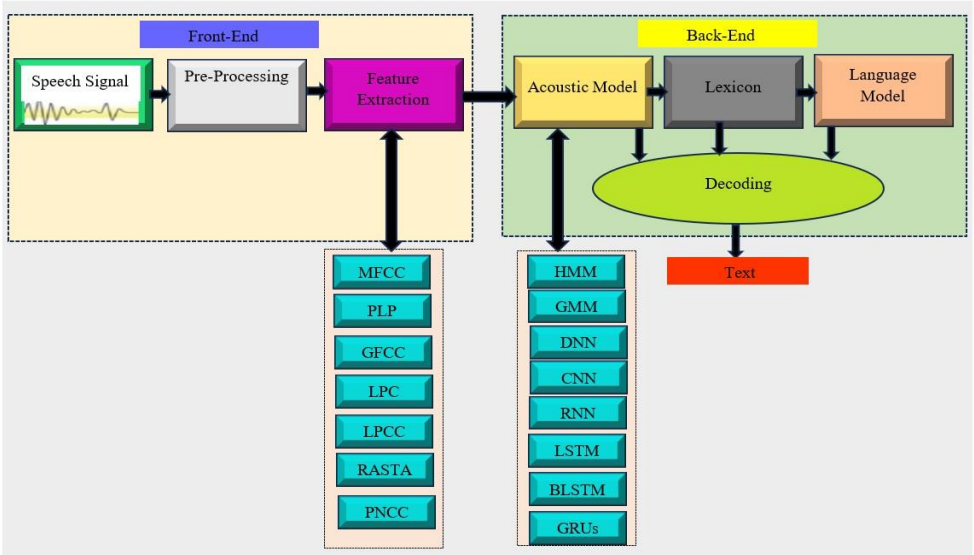


Figure 8: Combined architecture of feature extraction techniques with acoustic models

Neural network models containing 5 fully connected layers with 1024 hidden units in each layer, and the last layer, i.e., softmax layer having 42 output targets is used as DNN architecture. DNNs is pre-trained using the cross-entropy training and then Hessian Free sequence-training is applied for training purpose (Kingsbury, Sainath, & Soltau, 2012). The DNN-HMM system uses the same pre-trained DNN architecture (Tara N Sainath, Kingsbury, Mohamed, Saon, & Ramabhadran, 2014). Heteroscedastic Linear Discriminant Analysis (HLDA) is applied on softmax layer to reduce the dimensionality from 1024 to 42. By using the DNN-based acoustic model, Maximum Mutual Information (MMI) HMM training is applied. For MMI training, numerator and denominator lattice are used. The denominator

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

lattice represents the most likely word sequences for any training sentence and the numerator lattice represents the language model log probabilities.

The architecture given in (Tara N Sainath et al., 2015) is used as our baseline CNN architecture where CNN act as acoustic model. This baseline architecture consists of 2 convolutional layers with 512, 256 feature maps respectively and 4 fully connected layers of 1024 hidden units each. Each convolutional layer is followed by max pooling layer having size 2×1. All the neurons are maxout except for the last layer.

The input to the RNN acoustic model is a sequence of feature vectors extracted from the audio signal. RNN consisting of 3 recurrent layers process the input sequence and captures temporal dependencies. Each recurrent layer maintains a hidden state that is updated at each time step based on the current input and the previous hidden state. Output of the recurrent layers is fed into 2 fully connected layers to generate predictions. In final layer, softmax activation function is applied to produce a probability distribution over the output classes.

LSTM acoustic model used the following configuration for experimental purposes. 5 LSTM layers are stacked to learn complex temporal patterns. Each LSTM layer consists of 256 LSTM units (cells). After the LSTM layers, two fully connected layers transform the LSTM outputs into the desired output space. These layers have Rectified Linear Unit (ReLU) activation functions. The final output layer is usually a softmax layer that provides probability distributions over the set of possible phonemes or sub-word units. 42 units in the softmax layer correspond to the number of target classes (e.g., phonemes, characters).

BLSTM networks used in ASR are bidirectional, meaning they have two parallel LSTM layers for each time step: one processing the sequence from start to end and another from end to start. This setup allows the network to have context from both past and future frames, improving recognition accuracy. 10 LSTM layers (5 Forward and 5 Backward) are stacked to learn complex temporal patterns. Each LSTM layer consists of 256 LSTM units. After the LSTM layers, 2 linear layers are applied to transform the LSTM outputs into the desired output space. These layers have ReLU activation functions. The final output layer is usually a softmax layer that provides probability distributions over the set of possible phonemes or sub-word units. 42 units in the softmax layer correspond to the number of target classes (e.g., phonemes, characters).

In GRU architecture, 4 GRU layers are stacked to capture the temporal dependencies of speech. The number of hidden units per layer used is 512 to significantly impacts the model's capacity and complexity. More units allow for capturing complex temporal patterns but require more training data and computational resources. Bidirectional GRUs analyze the speech sequence in both directions, potentially improving accuracy therefore GRUs are adjusted in bidirectional. A fully connected layer with 1024 units are applied for linear mapping. The final layer depends on the specific task within ASR. The output layer uses a softmax activation to predict the probability distribution over all possible phonemes. 42 units are used in softmax layer.

Hardware Details

The various ASR architecture designs are tested on a supercomputer named PARAM Shavak Yuva-II. It consists of 2 multicore CPUs having 18 cores each along with two number of accelerator cards (i.e., GPGPU). The computer has 64GB RAM, 8TB storage, Nvidia Pascal architecture-based co-processing technologies, and deep learning GPU software environment. It works under Ubuntu v22.04 operating system. Kaldi toolkit is used for the implementation with Python. Note that same hardware detail is used for all experiments performed in this paper.

Acoustic Modeling and Training

The different ASR systems use different acoustic models which have been trained using discriminative techniques. 39-dimensional features are generated from the speech. The speech signal is corrupted artificially with 20dB SNR using noizeus dataset to design the noisy dataset. The model is trained using clean dataset and tested for clean and noisy datasets respectively.

Training Criteria

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Asynchronous Stochastic Gradient Descent (ASGD) optimization strategy (Dean et al., 2012) is used to train the neural network with the CE criteria with a context window of 15 frames. Typically, CE is used as the objective function, and optimization is performed through ASGD

Experimental Results and Analysis

This section represents the performance of the various ASR which is analyzed in terms of WER (%). Table 4 shows the WER (%) obtained by the different ASR systems in clean environment.

Table 4: Result of ASR systems for different combinations of feature extraction techniques and acoustic models in clean environment in terms of WER (%).

Feature Extraction/ Acoustic Models	HMM	GMM	DNN	CNN	RNN	LSTM	BLSTM	GRU
MFCC	**13.1**	**12.7**	**12.5**	**12.2**	**12.3**	**12.1**	**12**	**12.3**
PLP	13.7	13.4	13.2	12.8	13.0	12.8	12.6	12.9
GFCC	13.4	13.1	12.9	12.5	12.8	12.6	12.4	12.7
LPC	13.8	13.5	13.3	12.9	13.1	12.9	12.7	13.0
LPCC	13.6	13.3	13.1	12.7	13.0	12.8	12.6	12.9
RASTA-PLP	13.5	13.2	13.0	12.7	12.8	12.7	12.4	12.7
PNCC	13.3	12.9	12.7	12.3	12.6	12.4	12.2	12.5

From the analysis of the results presented in the above table, it is concluded that MFCC features performed well with BLSTM, outperforming all the feature extraction techniques in a clean environment. The reason for the best performance of the MFCC feature extraction technique is Mel filters are designed to mimic the human ear's response to different frequencies. It is based on the perceptual scale of pitches, which aligns the feature extraction process with how humans naturally process speech sounds. Moreover, its cepstral transformation also reduces channel variations' impact, such as different microphones or recording conditions. On the other hand, other feature extraction techniques are missing these characteristics, so they are lagging in recognition rate when compared with MFCC features. Table 5 represents the result achieved by the different combinations of feature extraction techniques with acoustic models in noisy environments in terms of WER (%).

Table 5: Result of ASR systems for different combinations of feature extraction techniques and acoustic models in noisy environments in terms of WER (%)

Feature Extraction/ Acoustic Models	HMM	GMM	DNN	CNN	RNN	LSTM	BLSTM	GRU
MFCC	16.1	15.7	15.5	15.2	15.3	15.1	15.0	15.3
PLP	16.7	16.4	16.2	15.8	16.0	15.8	15.6	15.9
GFCC	**15.8**	**15.5**	**15.2**	**15.0**	**15.0**	**15.0**	**14.8**	**14.9**
LPC	16.8	16.5	16.3	15.9	16.1	15.9	15.7	16.0
LPCC	16.6	16.3	16.1	15.7	16.0	15.8	15.6	15.9
RASTA-PLP	16.4	16.1	15.9	15.6	15.7	15.6	15.3	15.5
PNCC	15.3	15.9	15.7	15.3	15.6	15.4	15.2	15.5

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The results given in Table 5 show that GFCC features offer the best recognition rate with the BLSTM acoustic model in noisy environments over all the feature extraction techniques. The reason for GFCC performance is GFCCs use the Gammatone filterbank, which mimics the filtering characteristics of the human auditory system. Gammatone filterbank models the cochlear filtering process to effectively isolate important features from the signal by suppressing noise. GFCC also offers better frequency resolution at low frequencies than other feature extraction techniques, making it best in noisy environments.

This review paper gives a technical overview of the different feature extraction and acoustic modeling approaches widely used nowadays for ASR. An ASR system has mainly three stages: feature extraction stage, classification stage, and language modeling. Various feature extraction methods have been proposed, having different characteristics, and performing well in different scenarios. As discussed in the classification stage, the approach that transformed ASR research was the HMMs. GMM-HMM models require significant training data and can struggle with complex acoustic environments compared to DNN-HMM models. While DNN-HMM has become the dominant approach, understanding GMM-HMM provides a foundational knowledge of acoustic modeling in ASR. Although considerable accuracies were obtained from ASR systems based on HMMs, these are still far from achieving an optimal ASR system by themselves. Hence, numerous acoustic models, either based on the concept of integrating HMMs with another approach or direct modeling have been proposed. However, in recent years, CNNs have also been adopted in ASR systems, where numerous researchers proved the superiority of CNNs over ANNs, but most importantly researchers also showed that CNNs can achieve better results than HMMs. BLSTM has shown its capability in ASR due to bidirectional temporal

CONCLUSION

ASR has achieved new heights over the past few decades, due to advancements in machine learning and deep learning. The availability of large-scale annotated datasets also helped the Punjabi ASR research. Feature extraction techniques, such as MFCCs and GFCCs, play a crucial role in the performance of ASR systems. MFCCs have been widely adopted due to their efficiency and effectiveness in capturing essential audio features, whereas GFCCs offer superior performance in noisy environments by better modeling the human auditory system. The transition from traditional techniques like HMMs and GMMs to more sophisticated DNNs, CNNs, and RNNs has significantly improved the accuracy and robustness of ASR systems. However, bidirectional temporal modeling by BLSTM as acoustic model, made it the best acoustic model. Two research questions were discussed in the introduction section, for first RQ answer is MFCC and GFCC are the best performing feature extraction techniques in clean and noisy environments respectively. Second RQ answer is BLSTM. Throughout the experiments if we see Table 4 and Table 5 results, BLSTM performed well with every feature extraction technique due to its bidirectional temporal modeling capability.

Despite these advancements, challenges remain in achieving human-level performance in diverse and complex acoustic environments. Issues such as background noise, speaker variability, accents, and language differences continue to pose significant hurdles. To address these challenges, enhancing noise robustness, dialects, pitch, and tonal characteristics will be focused in future.

REFRENCES

[1] Baum, L. E., & Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society, 73*(3), 360-363.
[2] Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics, 41*(1), 164-171.
[3] Chen, B., Xing, L., Liang, J., Zheng, N., & Príncipe, J. C. (2014). Steady-State Mean-Square Error Analysis for Adaptive Filtering under the Maximum Correntropy Criterion. *IEEE signal processing letters, 21*(7), 880-884. doi:10.1109/LSP.2014.2319308
[4] Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(1), 30-42.
[5] Davis, S. B., & Mermelstein, P. (1990). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *Readings in speech recognition* (pp. 65-74): Elsevier.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[6] Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., . . . Le, Q. V. (2012). *Large scale distributed deep networks*. Paper presented at the Advances in neural information processing systems.
[7] Graves, A., Mohamed, A.-r., & Hinton, G. (2013, 26-31 May 2013). *Speech recognition with deep recurrent neural networks*. Paper presented at the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing.
[8] Guglani, J., & Mishra, A. N. (2020). Automatic speech recognition system with pitch dependent features for Punjabi language on KALDI toolkit. *Applied Acoustics, 167*, 107386. doi:https://doi.org/10.1016/j.apacoust.2020.107386
[9] Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America, 87*(4), 1738-1752.
[10] Hermansky, H., & Morgan, N. (1994). RASTA processing of speech. *IEEE transactions on speech and audio processing, 2*(4), 578-589.
[11] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., . . . Sainath, T. N. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine, 29*(6), 82-97.
[12] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.
[13] Kim, C., & Stern, R. M. (2016). Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing, 24*(7), 1315-1329.
[14] Kingsbury, B., Sainath, T. N., & Soltau, H. (2012). *Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization*. Paper presented at the Interspeech.
[15] Koehler, J., Morgan, N., Hermansky, H., Hirsch, H.-G., & Tong, G. (1994). *Integrating RASTA-PLP into speech recognition*. Paper presented at the Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on.
[16] Li, X., & Wu, X. (2015). *Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition*. Paper presented at the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
[17] Mohamed, A.-r., Sainath, T. N., Dahl, G., Ramabhadran, B., Hinton, G. E., & Picheny, M. A. (2011). *Deep Belief Networks using discriminative features for phone recognition*. Paper presented at the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). http://dx.doi.org/10.1109/icassp.2011.5947494
[18] Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47*(1), 90-100. doi:https://doi.org/10.1016/S0022-2496(02)00028-7
[19] O'Shaughnessy, D. (1988). Linear predictive coding. *IEEE potentials, 7*(1), 29-32.
[20] Palaz, D., Magimai-Doss, M., & Collobert, R. (2019). End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition. *Speech communication, 108*, 15-32.
[21] Pasricha, V., & Aggarwal, R. (2016). *Hybrid architecture for robust speech recognition system*. Paper presented at the 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE).
[22] Passricha, V., & Aggarwal Rajesh, K. (2019). A Hybrid of Deep CNN and Bidirectional LSTM for Automatic Speech Recognition. In *Journal of Intelligent Systems* (Vol. 0).
[23] Passricha, V., & Aggarwal, R. K. (2019). End-to-End Acoustic Modeling Using Convolutional Neural Networks. In *Intelligent Speech Signal Processing* (pp. 5-37): Elsevier.
[24] Passricha, V., & Aggarwal, R. K. (2020). A comparative analysis of pooling strategies for convolutional neural network based Hindi ASR. *Journal of Ambient Intelligence and Humanized Computing*. doi:10.1007/s12652-019-01325-y
[25] Quen-Zong, W., Jou, I. C., & Suh-Yin, L. (1997). On-line signature verification using LPC cepstrum and neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 27*(1), 148-153. doi:10.1109/3477.552197
[26] Ravanelli, M., Brakel, P., Omologo, M., & Bengio, Y. (2018). Light Gated Recurrent Units for Speech Recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence, 2*(2), 92-102. doi:10.1109/TETCI.2017.2762739
[27] Ren, J. S., & Xu, L. (2015). *On Vectorization of Deep Convolutional Neural Networks for Vision Tasks*. Paper presented at the Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, Texas.
[28] Robinson, T., Hochberg, M., & Renals, S. (1996). The use of recurrent neural networks in continuous speech recognition. In *Automatic speech and speaker recognition* (pp. 233-258): Springer.
[29] Sainath, T. N., Kingsbury, B., Mohamed, A.-r., Saon, G., & Ramabhadran, B. (2014). *Improvements to filterbank and delta learning within a deep neural network framework*. Paper presented at the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Research Article

[30] Sainath, T. N., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A.-r., Dahl, G., & Ramabhadran, B. (2015). Deep convolutional neural networks for large-scale speech tasks. *Neural networks, 64*, 39-48.
[31] Sainath, T. N., Kingsbury, B., Soltau, H., & Ramabhadran, B. (2013). Optimization Techniques to Improve Training Speed of Deep Neural Networks for Large Speech Tasks. *IEEE Transactions on Audio, Speech, and Language Processing, 21*(11), 2267-2276. doi:10.1109/TASL.2013.2284378
[32] Sak, H., Senior, A., & Beaufays, F. (2014). *Long short-term memory recurrent neural network architectures for large scale acoustic modeling*. Paper presented at the Interspeech.
[33] Viikki, O., & Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech communication, 25*(1-3), 133-147.
[34] Yu, D., Seide, F., & Li, G. (2012). *Conversational Speech Transcription Using Context-Dependent Deep Neural Networks*. Paper presented at the ICML.
[35] Zhao, X., & Wang, D. (2013). *Analyzing noise robustness of MFCC and GFCC features in speaker identification*. Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.